Saliency Detection by Multi-Task Sparsity Pursuit

Congyan Lang, Guangcan Liu, *Member, IEEE*, Jian Yu, and Shuicheng Yan, *Senior Member, IEEE*,

***Abstract*—This paper addresses the problem of detecting salient areas within natural images. We shall mainly study the problem under unsupervised setting, namely saliency detection without learning from labeled images. A solution of multi-task sparsity pursuit is proposed to integrate multiple types of features for detecting saliency collaboratively. Given an image described by multiple features, its saliency map is inferred by seeking the consistently sparse elements from the joint decompositions of multiple feature matrices into pairs of low-rank and sparse matrices. The inference process is formulated as a constrained nuclear norm and $\ell_{2,1}$-norm minimization problem, which is convex and can be solved efficiently with augmented Lagrange multiplier method. Compared to previous methods, which usually make use of multiple features by combining the saliency maps obtained from individual features, the proposed method seamlessly integrates multiple features to jointly produce the saliency map with a single inference step, and thus produces more accurate and reliable results. Besides the unsupervised setting, the proposed method can be also generalized to incorporate the top-down priors obtained from supervised environment. Extensive experiments well validate its superiority over other state-of-the-art methods.**

***Index Terms*—saliency detection, multi-feature modeling, multi-task learning, sparse and low-rank.**

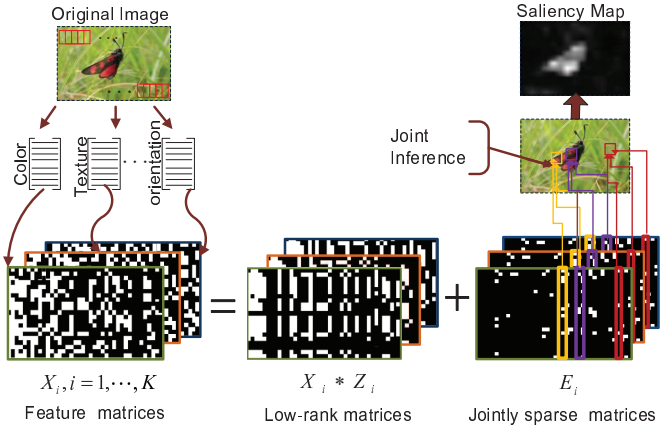


Fig. 1. For a given image, first, we extract K types of features, resulting in K types of feature matrices $X_1, X_2, \cdots, X_K$, with each X_i corresponding to a certain type of feature. Second, its saliency map is inferred by seeking the consistently sparse elements E from the joint decompositions of multiple feature matrices X_i into pairs of low-rank and sparse matrices. Note here that our method can also handle the saliency detection problem based on a single feature (i.e., $K = 1$).

I. Introduction

VISUAL attention is crucial in determining visual experience, leading to the challenging problem of *saliency detection* that is an important function for image processing and understanding [1], [2], [3], [4]. Saliency detection is related to many applications, such as automatic image cropping [5], image thumbnailing [6], image/video compressing [7] and image collection browsing [8]. Therefore saliency detection problem has been extensively studied in signal processing, computer vision, machine learning and even biological literature (e.g., [9], [10], [11], [12], [13], [14], [15], [16], [17], [18]). According to whether the detection procedure requires human interaction or not, existing methods are divided into two categories: bottom-up (unsupervised) and top-down (supervised). In this paper, for ease of presentation, we shall firstly study the problem under the first setting, namely no learning process from labeled images is taken into account. Then, it will be shown that the proposed model can be naturally generalized to incorporate the top-down priors obtained from supervised environment.

Saliency detection is to automatically select the sensory information that is notable to human vision system. From the perspective of computer vision, the goal is to find the image regions where one or more of their features differ from those in the surroundings. As a comprehensive task, it contains several issues, such as how to extract effective features and what is the optimal criterion for measuring saliency. Through lots of efforts, researchers have found several effective feature descriptors, mainly including color, texture and orientation, as surveyed in [17]. For a certain feature schema, many computational methods have been established for measuring and detecting saliency (e.g., [2], [6], [19], [20]). However, the salient regions can seldom be well described by only a single feature, and generally it is hard for such methods to handle well a wide range of images. This is because a single feature descriptor usually only captures one aspect of the visual information. For example, the color based descriptors may not handle well the images with rich textures. Therefore in this paper we study a basic problem as follows:

- Provided that each image is described by several types of features, how to integrate these multiple features for accurate and reliable saliency detection?

In fact, it is generally accepted that saliency detection may benefit from the integration of multiple visual features. Unfortunately, most existing literature on this direction focuses

Congyan Lang is with the Department of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, 100044, e-mail: cylang@bjtu.edu.cn.

Guangcan Liu is with the Department of Electrical and Computer Engineering, National University of Singapore, 117576, Singapore, e-mail: eleliug@nus.edu.sg.

Jian Yu is with the Department of Computer and Information Technology, Beijing Jiaotong University, Beijing, China, 100044, e-mail: jianyu@bjtu.edu.cn

Shuicheng Yan is with the Department of Electrical and Computer Engineering, National University of Singapore, 117576, Singapore, e-mail: eleyans@nus.edu.sg

Manuscript received April 16 , 2011; revised September 12, 2011.

on the "naive" combination frameworks. Typically, after the saliency maps are computed for each of the features *individually*, they are normalized and then combined in a linear or non-linear fashion for producing a final saliency map [2], [21], [22], [23]. The cross-feature information is not well utilized in the inference process and it is often difficult for such naive approaches to produce reliable results.

To make effective use of multiple features, in this paper we propose a *multi-task sparsity pursuit* (MTSP) model for saliency detection. Figure 1 outlines the proposed method, which differs significantly from the previous methods in its motivation and methodology. We treat saliency detection as a sparsity pursuit problem and integrate multiple types of features for detecting saliency collaboratively. Since the cross-feature information has been well considered, such a joint inference schema can produce more accurate and reliable results than the models of producing the saliency maps individually. The inference process is formulated as a constrained nuclear norm and $\ell_{2,1}$-norm minimization problem, which is convex and can be solved efficiently with augmented Lagrange multiplier (ALM) [24] method. Besides the ability of modeling multiple features, as will be seen, another advantage of MTSP is that it can be naturally generalized to incorporate the top-down priors so as to produce more accurate results. In summary, the contributions of this work mainly include:

1. We propose a sparsity pursuit framework for saliency detection. Compared to existing models, the proposed framework seamlessly integrates multiple types of features into a unified inference procedure, which is formulated as a convex optimization problem. With some mild modifications, the proposed model can also handle the top-down priors from supervised environment.
2. Based on the proposed framework, we establish effective algorithms for saliency detection. Experimental results show that our algorithms remarkably outperform other state-of-the-art algorithms. Our algorithms are also computationally efficient.
3. The proposed MTSP is a general multi-task method for achieving sparsity jointly. It may be useful for other related problems.

The remainder of this paper is organized as follows. We discuss and analyze the related work in Section II. We propose in Section III the multi-task sparsity pursuit algorithm and its extension for saliency detection. Experimental results are given in Section IV. Finally, we give the concluding remarks in Section V.

II. Related Work

The major difference among different saliency detection models is the mechanism to measure saliency. Many of the saliency approaches [2], [18], [25] construct biologically-plausible mechanisms based on the findings from psychology or neurobiology [26]. As one of the earliest methods, Itti et al. [2] proposed a center-surround operation as local feature contrast in color, intensity and orientation of an image. Le Meur et al. [25], [27] modeled the bottom-up visual attention using a coherent computational approach. Psycho-visual space is used to combine the visual features and the saliency is computed using a contrast sensitivity function. The saliency is also computed by a discriminant center-surround hypothesis using mutual information [13]. The saliency value at each location is essentially explained as the local contrast in one feature or more. There lacks of global measurement and so these methods may not produce satisfactory results, as pointed out by [10]: "Although this approach may be biologically plausible, it is suboptimal for computer vision".

More recently, the vast majority of saliency detection methods attempt to detect saliency based on more mathematically motivated principles. Hou et al. [28] computed visual saliency by extracting spectral residuals in the amplitude spectrum of Fourier transform. The recent trends on estimating saliency are emphasizing on utilizing graph model [19], maximum information sampling [11] and subspace analysis [20]. These methods suffer from the limitation that cluttered backgrounds may yield higher saliency as such backgrounds possess high global exception in the cases with complex scenes. Meanwhile, the object borders are often assigned with high saliency than the salient regions, even if the neighborhood size parameter is well tuned. In order to overcome these issues, several methods [1], [10] incorporated the image segmentation or pixel clustering into saliency detection. However, the good performance heavily relies on the quality of image segmentation, which itself is a challenging problem.

As the salient target is usually small which implies sparsity, saliency detection can benefit from sparse signal analysis techniques. In particular, the recently established Robust Principal Component Analysis (RPCA) [29] and its variations may fit well to the saliency detection problem [30]. Given a matrix X_0, RPCA aims at decomposing it into a low-rank matrix A_0 and a sparse one E_0 by minimizing

$$\begin{aligned} \min_{A_0,E_0} \quad & \|A_0\|_* + \lambda\|E_0\|_1, \\ \text{s.t.} \quad & X = A_0 + E_0, \end{aligned}$$

where $\|\cdot\|_*$ denotes the matrix *nuclear norm* (sum of the singular values of a matrix) [31], which is a convex relaxation of the rank function, $\|\cdot\|_1$ is the ℓ_1*-norm* and the parameter $\lambda > 0$ is used to balance the effects of the two parts. The technique used in this paper is a variation of RPCA, with better connections to the problem space of saliency detection.

As a pattern analysis problem, saliency detection performance heavily depends on the choice of feature space. What is more, it is hard to find a single feature descriptor that can generally work well for various images with diverse properties. So, the importance of combining multiple features has been widely discussed (e.g., [2], [21], [22], [23]). However, as discussed in Section I, existing methods may not fully capture the advantages of multiple features, because the combination is performed on the saliency maps inferred individually from each feature. The cross-feature information is not well utilized in the inference process and thus it is often difficult for these methods to produce generally reliable results. Our MTSP method provides an effective solution for this issue: the feature fusion is performed during the saliency map inference process such that the cross-feature information is well utilized.

The prior knowledge on scene context or specific objects is also helpful for identifying saliency regions, leading to the exploration of combining bottom-up and top-down information [17], [22], [23], [32] for saliency detection. A straightforward way is to integrate the bottom-up and top-down components obtained individually, as done in [33]. Since the inference process does not benefit from the integration, this simple approach may not produce satisfactory results. Therefore, most researchers choose to modify their bottom-up models for including available top-down priors [34], [35]. The proposed MTSP method, as will be seen, can be naturally generalized to incorporate some top-down priors.

III. Multi-task Sparsity Pursuit for Saliency Detection

In this section, we elaborate on the proposed MTSP model. For easy of presentation, we shall focus on the unsupervised setting, i.e., bottom-up saliency detection. The generalization for handling the top-down priors will be discussed at the end of this section.

A. Problem Formulation

For efficiency, we use image patches other than pixels as basic image elements. Namely, for a given image, we partition it into non-overlapping patches of size $p \times q$ pixels. Then a pixel is salient if and only if the patch containing this pixel is judged to be salient. By choosing an appropriate feature descriptor to describe each patch, saliency detection problem can be formulated as follows.

Formulation 1: Let $X = [x_1, x_2, \cdots, x_N]$ with size $d \times N$ be a feature matrix, each column of which is a feature vector x_i corresponding to an image patch P_i. Then the task is to find an assignment function $S(P_i) \in [0, 1]$. The function $S(P_i)$ is referred to as saliency map, where the higher value indicates higher salient location.

A weak point of the above formulation is that only one type of feature is considered. For better performance, we consider the following multiple features based problem formulation.

Formulation 2: Let $X_1, X_2, \cdots, X_K$ be K feature matrices for K types of features, where the columns in different matrices with the same index correspond to the same image patch. The size of each X_i is $d_i \times N$, where d_i is the feature dimension and N is the number of patches. Then the task is to find an assignment function $S(P_i) \in [0, 1]$ by integrating the feature matrices $X_1, \cdots, X_K$.

B. Multi-Task Sparsity Pursuit

For better understanding, we explore Formulation 1 at first, then we shall establish an algorithm for the multi-feature case (Formulation 2) accordingly.

1) Single-Feature Case (Formulation 1): The task described by Formulation 1 is to find a criterion for measuring and detecting saliency. In human vision system, usually, only the *distinctive* sensory information is selected for further processing. From this perspective, the salient targets should be different from the background (non-salient) patches. Moreover, there usually exists strong correlation among the background patches, i.e., the background patches are usually *self-represented*. This analysis suggests that the matrix X may be decomposed into a salient part and a non-salient part as follows:

$$X = XZ_0 + E_0, \tag{1}$$

where XZ_0 denotes the non-salient part which can be reconstructed by itself, Z_0 denotes the reconstruction coefficients, and E_0 denotes the rest part ($E_0 = X - XZ_0$) corresponding to the salient targets.

Without imposing any restrictions, there are infinite number of solutions (with respect to Z_0 and E_0) to (1). To seek a solution that is useful for saliency detection, we need some criteria for characterizing the matrices Z_0 and E_0. To this end, we have two basic principles. On one hand, as adopted by most approaches in computer vision (e.g., [2], [10]), we assume that only a small fraction of patches are salient, i.e., the matrix E_0 is *sparse*. The connection between sparsity and saliency is also consistent with the fact that only a small subset of sensory information is selected for further processing in human vision system. On the other hand, the strong correlation among the background patches suggests that the matrix Z_0 may have the property of *low-rankness*. In summary, for a matrix $X = [x_1, x_2, \cdots, x_N]$ with each x_i representing the i-th patch, it is appropriate to infer the salient patches by solving the following low-rank representation (LRR) [36] problem:

$$\begin{aligned} \min_{Z_0, E_0} \quad & \|Z_0\|_* + \lambda \|E_0\|_{2,1}, \\ \text{s.t.} \quad & X = XZ_0 + E_0, \end{aligned} \tag{2}$$

where $\|\cdot\|_*$ denotes the matrix *nuclear norm* (sum of the singular values of a matrix) [31], which is a convex relaxation of the rank function [29], the parameter $\lambda > 0$ is used to balance the effects of the two parts, and $\|\cdot\|_{2,1}$ is the $\ell_{2,1}$-*norm* [36] defined as the sum of ℓ_2 norms of the columns of a matrix:

$$\|E_0\|_{2,1} = \sum_i \sqrt{\sum_j (E_0(j,i))^2}.$$

Here $E_0(j, i)$ is the (j, i)-th entry of E_0. Since the minimization of $\ell_{2,1}$-norm encourages the columns of E_0 to be zero (i.e., have sparse columns), it fits well our saliency detection problem. For a column corresponding to the i-th patch, larger (smaller) magnitude implies that the patch is more salient (non-salient), i.e., the sparse matrix E_0 naturally measures visual saliency.

Let E_0^* be the optimal solution (with respect to E_0) to problem (2). To obtain a saliency score for the i-th patch P_i, we only need a simple post-processing step to quantify the

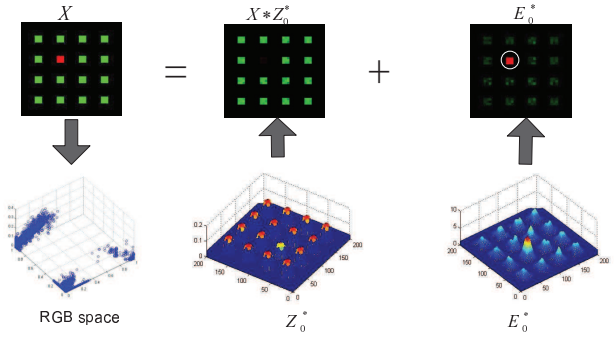


Fig. 2. An example of using LRR to perform saliency detection. For the given image represented by the feature matrix X, it can be seen that the non-salient and salient targets are naturally identified by XZ_0^* and E_0^*, respectively. Here, Z_0^* and E_0^* are obtained by solving problem (2).

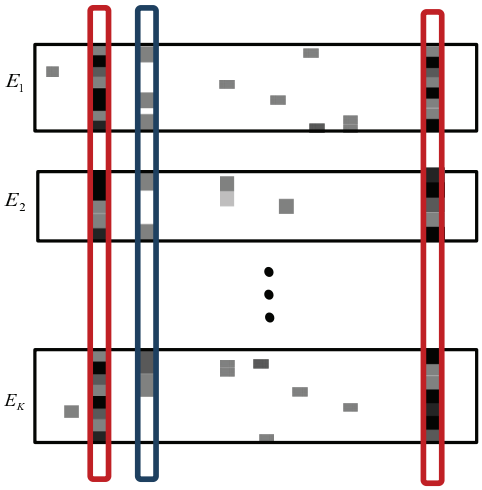


Fig. 3. Illustrating the minimization of the $\ell_{2,1}$-norm defined on E. Generally, this technique is to enforce the entries of the matrices $E_1, E_2, \cdots, E_K$ to have jointly consistent columns. Since the columns in different matrices with the same index correspond to the same image patch, this technique is also to encourage different features to produce consistent saliency maps. In this way, the cross-feature information is naturally modeled such that the multiple features will take effects collaboratively.

response of the sparse matrix:

$$\mathrm{S}(P_i) = \|E_0^*(:,i)\|_2 = \sqrt{\sum_j (E_0^*(j,i))^2}, \tag{3}$$

where $\|E_0^*(:,i)\|_2$ denotes the ℓ_2-norm of the i-th column of E_0^* and $E_0^*(j,i)$ is its (j,i)-th entry. The large score of $\mathrm{S}(P_i)$ means that the patch P_i has high probability to be salient. In this way, the task of saliency detection is performed by solving the LRR problem (2). Figure 2 exemplifies how the salient targets are found.

2) Multi-Feature Case (Formulation 2): The above LRR can only model a certain type of visual feature, which cannot be directly used for multi-feature cases. To combine together multiple features, as adopted by existing methods (e.g., [2]), an intuitive approach is to directly combine the saliency maps obtained from individual features. However, the inference of the individual saliency map does not well utilize the cross-feature information, and thus it is often difficult to produce accurate and reliable results. Here, we propose a new solution of multi-task sparsity pursuit (MTSP), which is a multi-task generalization of LRR. MTSP seeks a jointly sparse matrix E by solving the following convex optimization problem:

$$\begin{aligned} \min_{\substack{Z_1,\cdots,Z_K \\ E_1,\cdots,E_K}} \quad & \sum_{i=1}^{K} \|Z_i\|_* + \lambda \|E\|_{2,1} \\ \text{s.t.} \quad & X_i = X_i Z_i + E_i, i = 1, \cdots, K, \end{aligned} \tag{4}$$

where $E = [E_1; E_2; \cdots; E_K]$ is formed by vertically concatenating $E_1, E_2, \cdots, E_K$ together along column. The integration of multiple features is "seamlessly" performed by minimizing the $\ell_{2,1}$-norm of E. Namely, this technique will enforce the columns of $E_1, E_2, \cdots, E_K$ to have jointly consistent magnitudes, i.e., they are all large or they are all small, as shown in Figure 3.

Let $\{E_1^*, \cdots, E_K^*\}$ be the optimal solution (with respect to E_i's) to problem (4). To obtain a saliency score for the i-th patch P_i, similar as the single-feature case, we quantify the response of the sparse matrices as follows:

$$\mathrm{S}(P_i) = \sum_{j=1}^{K} \|E_j^*(:,i)\|_2 \tag{5}$$

where $\|E_j^*(:,i)\|_2$ denotes the ℓ_2-norm of the i-th column of E_j^*. The large score of $\mathrm{S}(P_i)$ means that the patch P_i has high probability to be salient. While $K = 1$ (i.e., single-feature case), it can be seen that the formulation (4) falls back to (2). The saliency function defined by (5) is also a generalization of (3). Hence, the proposed MTSP model can actually handle both cases for single-feature and multi-feature. Algorithm 1 summarizes the whole procedure of MTSP based saliency detection.

Algorithm 1 Saliency Detection by MTSP

Input: An image and the required parameters.

1. Compute K feature matrices by extracting k types of features to describe each image patch.

2. Obtain the sparsity matrices $E_1, E_2, \cdots, E_K$ by solving problem (4).

3. Compute the saliency map by (5).

Output: A map that encodes the saliency value of each image patch.

C. Optimization Procedure

Problem (4) is convex and can be optimized efficiently. We first convert it into the following equivalent problem:

$$\begin{aligned} \min_{\substack{J_1,\cdots,J_K \\ Z_1,\cdots,Z_K \\ E_1,\cdots,E_K}} \quad & \sum_{i=1}^{K} \|J_i\|_* + \lambda \|E\|_{2,1} \\ \text{s.t.} \quad & X_i = X_i Z_i + E_i, \\ & Z_i = J_i, i = 1, \cdots, K. \end{aligned} \tag{6}$$

This problem can be solved with the augmented Lagrange multiplier (ALM) method [24], which minimizes the following

Algorithm 2 Solving Problem (4) by Inexact ALM

Input: Data matrices $\{X_i\}$, parameter λ.

while not converged **do**

1. Fix the others and update $J_1, \cdots, J_K$ by

$$J_i = \arg\min_J \frac{1}{\mu}||J||_* + \frac{1}{2}||J_i - (Z_i + \frac{W_i}{\mu})||_F^2.$$

2. Fix the others and update $Z_1, \cdots, Z_K$ by

$$Z_i = M(X_i^T(X_i - E_i) + J_i + \frac{X_i^T Y_i - W_i}{\mu}),$$

where $M = (I + \sum_{i=1}^K X_i^T X_i)^{-1}$.

3. Fix the others and update $E = [E_1; E_2; \cdots; E_K]$ by

$$E = \arg\min_E \frac{\lambda}{\mu}||E||_{2,1} + \frac{1}{2}||E - G||_F^2,$$

where G is formed by vertically concatenating the matrices $X_i - X_i Z_i + \frac{Y_i}{\mu}, i = 1, \cdots, K$ together along column.

4. Update the multipliers

$$\begin{aligned} Y_i &= Y_i + \mu(X_i - X_i Z_i - E_i), \\ W_i &= W_i + \mu(Z_i - J_i). \end{aligned}$$

5. Update the parameter μ by

$$\mu = \min(\rho\mu, 10^{10}),$$

where the parameter ρ takes the role of controlling the convergence speed. It is set as $\rho = 1.1$ in all experiments.

6. Check the convergence condition: $X_i - X_i Z_i - E_i \to 0$ and $Z_i - J_i \to 0$, $i = 1, \cdots, K$.

end while

Output: The optimal solution E^*.

augmented Lagrange function:

$$\mathcal{L} = \lambda\|E\|_{2,1} + \sum_{i=1}^K (\|J_i\|_* + \langle Y_i, X_i - X_i Z_i - E_i\rangle + \langle W_i, Z_i - J_i\rangle + \frac{\mu}{2}\|X_i - X_i Z_i - E_i\|_F^2 + \frac{\mu}{2}\|Z_i - J_i\|_F^2),$$

where $Y_1, \cdots, Y_K$ and $W_1, \cdots, W_K$ are Lagrange multipliers, and $\mu > 0$ is a penalty parameter. The inexact ALM method, also called alternating direction method, is outlined in Algorithm 2. Notice that the sub-problems of the algorithm are convex and they all have closed-form solutions. Step 1 is solved via the singular value thresholding operator [37], while Step 3 is solved via Lemma 3.2 of [36].

1) On the Convergence Properties: When the objective function is smooth, the convergence of the exact ALM algorithm has been proven in [38]. For inexact ALM, which is a variation of exact ALM and also called as alternating direction method (ADM), its convergence has also been well studied when the number of blocks is not more than two [24], [39]. Up to present, it is still difficult to ensure the convergence of inexact ALM with three or more blocks [40]. Since there are $2K + 1$ ($K = 3$ in this work) blocks in Algorithm 2 and the objective function in (4) is not smooth, it would be difficult to *strictly* prove the convergence in theory.

Fortunately, there actually exist some guarantees for enuring the convergence of Algorithm 2. According to the theoretical results in [41], two conditions are *sufficient* (but may not necessary) for Algorithm 2 to converge: the first condition is that the feature matrices X_i ($i = 1, \cdots, K$) are of full column rank; the second one is that the optimality gap produced in each iteration step is monotonically decreasing, namely the error

$$\epsilon_l = \|(Z_1^l, \cdots, Z_K^l, J_1^l, \cdots, J_K^l) - \arg\min_{Z_i's, J_i's} \mathcal{L}\|_F^2$$

is monotonically decreasing, where Z_i^l (resp. J_i^l) denotes the solution produced at the l-th iteration, $\arg\min_{Z_i's, J_i's} \mathcal{L}$ indicates the "ideal" solution obtained by minimizing the Lagrange function $\mathcal{L}$ with respect to all $Z_1, \cdots, Z_K, J_1, \cdots, J_K$ simultaneously. The first condition is easy to obey, since Problem (4) can be converted into an equivalent problem where the full column rank condition is always satisfied (we will show this in the next subsection). For the monotonically decreasing condition, although it is difficult to *strictly* prove it, the convexity of the Lagrange function could guarantee its validity to some extend. So, it could be well expected that Algorithm 2 has good convergence properties. Moreover, inexact ALM is known to *generally* perform well in reality, as illustrated in [40].

2) Computational Complexity: Let the size of X_i be $d_i \times N$. Without loss of generality, suppose $d_1 = d_2 = \cdots = d_K = d$, then the computation complexity of Algorithm 2 is $O(N^3)$, which is inefficient when the number of image patches is large (i.e., the image is large). However, the complexity can be further reduced to $O(d^3 + d^2 N)$ (assume $d \leq N$) by utilizing the theories established by Liu et al. [42]. From Theorem 4.3 of [42], the optimal solution Z_i^* (with respect to the variable Z_i) always lies within the subspace spanned by the rows of X_i. So, Problem (4) can be equivalently converted into a simpler problem by replacing Z_i with $Q_i S_i$:

$$\begin{aligned} \min_{\substack{S_1,\cdots,S_K \\ E_1,\cdots,E_K}} \quad & \sum_{i=1}^K \|S_i\|_* + \lambda\|E\|_{2,1} \\ \text{s.t.} \quad & X_i = A_i S_i + E_i, i = 1, \cdots, K, \end{aligned} \tag{7}$$

where $A_i = X_i Q_i$ and Q_i is computed by orthogonalizing the columns of X_i^T. The above problem can be solved in a similar way as that for problem (4). As the number of the columns of A_i is at most d (assume $d \leq N$), the computational complexity is $O(d^3 + d^2 N)$, which is quite efficient because the feature dimension d is generally small compared with n. For example, $d \leq 13$ in our experiments. Thus computational complexity is reduced to $O(N)$, where N is the number of patches.

D. Generalized to Handle Top-down Priors

Saliency detection may also benefit from the labeled data, from which some kinds of object-specific or global-specific information can be inferred for identifying salient targets [32], [43], [44]. Up to present, MTSP has only considered the low-level visual features, i.e., bottom-up saliency detection without learning from labeled images. Fortunately, it is actually natural for MTSP to handle the top-down priors

represented by a label vector. In this subsection, we further describe the generalized MTSP (G-MTSP) to incorporate the top-down priors obtained from supervised environment. To be more precisely, G-MTSP is to address the saliency detection problem formulated as follows:

Formulation 3: Besides the feature matrices $X_1, \cdots, X_K$, suppose there is also a label vector $\Omega = (\pi_1, \cdots, \pi_N)$ that roughly assigns each patch P_i a probability $\pi_i \in [0, 1]$ of being salient. Then the task is to find an assignment function $S(P_i) \in [0, 1]$ by making use of both the visual features encoded by $X_1, \cdots, X_K$ and the top-down priors encoded in the label vector Ω.

The label vector Ω can be computed from existing top-down saliency detection or object detection algorithms. To handle the priors encoded by such a label vector, we only need to generalize the $\ell_{2,1}$-norm defined on E, namely a *weighted* $\ell_{2,1}$-norm as follows:

$$\|E\|_{2,1}^{\Phi} = \sum_{i=1}^{N} \phi_i \|E(:, i)\|_2, \tag{8}$$

where $\Phi = (\phi_1, \phi_2, \cdots, \phi_N)$ generally denotes a vector that assigns a weight ϕ_i for the i-th column of E, $i = 1, \cdots, K$. When $\phi_i \equiv 1, i = 1, \cdots, K$, the above weighted $\ell_{2,1}$-norm identifies the traditional $\ell_{2,1}$-norm. While minimizing the weighted $\ell_{2,1}$-norm, a larger weight means stronger penalty on the corresponding column, leading to smaller response on that column. So, the saliency detection task described in Formulation 3 may be handled by modifying our formulation (4):

$$\begin{aligned} \min_{\substack{Z_1, \cdots, Z_K \\ E_1, \cdots, E_K}} \quad & \sum_{i=1}^{K} \|Z_i\|_* + \lambda \|E\|_{2,1}^{1-\Omega} \\ \text{s.t.} \quad & X_i = X_i Z_i + E_i, i = 1, \cdots, K, \end{aligned} \tag{9}$$

where Ω is a given label vector. When $\Omega = 0$, the above formulation falls back to (4) and so (9) is a generalization of (4). Actually, the LRR formulation (2) is also a special case of (9) with $K = 1$ and $\Omega = 0$.

Since $\pi_i \in [0, 1], \forall i$, problem (9) is always convex. To solve problem (9), we only need to replace Step 3 of Algorithm 2 with

$$E = \arg\min_{E} \frac{\lambda}{\mu} \|E\|_{2,1}^{1-\Omega} + \frac{1}{2} \|E - G\|_F^2. \tag{10}$$

The solution to the above problem can be found by following the same way as Lemma 3.2 of [36]. First, notice that the above problem can be solved column-by-column:

$$E(:, i) = \arg\min_{e} \frac{\lambda(1 - \pi_i)}{\mu} \|e\|_2 + \frac{1}{2} \|e - G(:, i)\|_2^2,$$

where $E(:, i)$ is the i-th column of E, $i = 1, \cdots, N$. Then, it is simple to see that the solution is given by

$$E(:, i) = \begin{cases} \frac{\|G(:,i)\|_2 - \lambda(1-\pi_i)}{\|G(:,i)\|_2} G(:, i), & \text{if } \lambda < \frac{\|G(:,i)\|_2}{1-\pi_i}, \\ 0, & \text{otherwise.} \end{cases}$$

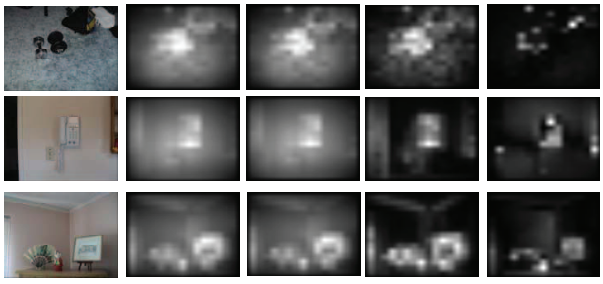


Fig. 4. Saliency maps obtained under different parameter settings on the Bruce dataset.

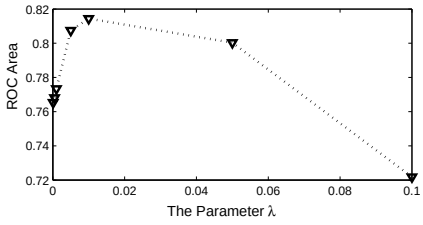


Fig. 5. The ROC area change curve as a function of parameter λ. There results are from the Bruce dataset.

Hence, problem (9) is solved by replacing Step 3 of Algorithm 2 with (10). The same as the analysis in Section III-C, the complexity of the optimization procedure is also $O(N)$ (assuming the feature dimension is small).

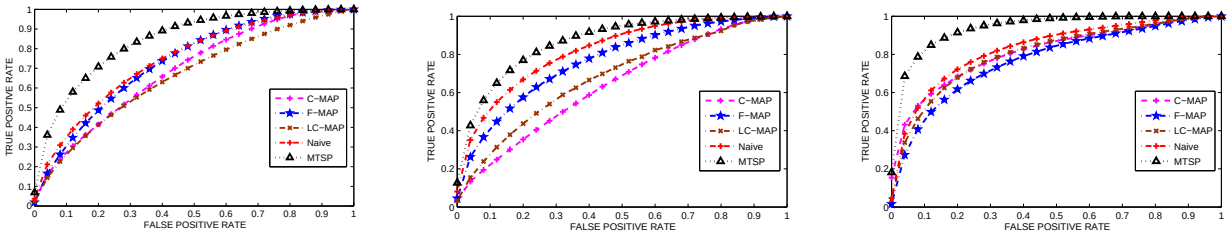
IV. Experiments and Results

A. Experimental Settings

1) Datasets: We tested the proposed algorithm on four public image datasets: MIT [17], Bruce [11], FiFa [34] and MSRA [32]. The MIT, Bruce and FiFa datasets consist of eye tracking data from several different users across images. All fixation patterns for a given image are added together for providing a spatial distribution of human fixation. The fourth dataset used in our experiments is the subset [9] of MSRA salient object dataset [32]. This dataset contains accurate object-contour based ground truth for quantitative evaluation.

2) Baselines: To show the advantages of the proposed MTSP model, we implemented eight state-of-the-art unsupervised models[1] for comparison, which are Itti model (IT) [2], graph based visual saliency (GBVS) [15], context aware based saliency detection (CSD) [14], self-information (SI) [11], sparse coding based method (SCSP) [30], saliency using natural statistics (SUN) [46] and other two methods using frequency domain features: spectral residual (SR) [28] and

[1]Source codes of these baseline algorithms are available at http://ilab.usc.edu/toolkit, http://www.klab.caltech.edu/∼harel/share/gbvs.php, http://bcmi.sjtu.edu.cn/∼houxiaodi, http://www−sop.inria.fr/members/Neil.Bruce/, http://cseweb.ucsd.edu/∼16zhang, http://webee.technion.ac.il/labs/cgm/Computer-Graphics-Multimedia/Software/, and http://ivrg.epfl.ch/supplementary_material/RK_CVPR09/.



(a) MIT Dataset (b) Bruce Dataset (c) MSRA Dataset

Fig. 6. To show the advantages of MTSP, we apply the proposed single task saliency detection by solving (2) for color, Local Energy [45], and Local Contrast [2] feature individually (denoted as "C-Map", "F-Map" and "LC-Map"). Also, we use a naive approach (denoted as "Naive") to combine multiple features by simply averaging the saliency maps obtained individually.

TABLE I
PERFORMANCE OF MTSP ON CONTAMINATED IMAGES (MSRA DATASET) WITH 5×5 PATCH NOISE (σ RANGES FROM 0.05 TO 0.45).

	σ	0.05	0.15	0.25	0.35	0.45
AUC	MTSP	**0.8922**	**0.8824**	**0.8565**	**0.8262**	**0.7898**
	C-MAP	0.8104	0.7892	0.7384	0.7155	0.6913
	F-MAP	0.7921	0.7759	0.6965	0.6046	0.5748
	LC-MAP	0.8092	0.7821	0.7719	0.7332	0.7037
	Naive	0.8318	0.8130	0.7850	0.7501	0.7175
CC	MTSP	**0.6232**	**0.6068**	**0.5536**	**0.5151**	**0.4831**
	C-MAP	0.6055	0.5419	0.4811	0.3945	0.3656
	F-MAP	0.5114	0.4175	0.2616	0.1528	0.1347
	LC-MAP	0.5763	0.5449	0.4552	0.4442	0.3801
	Naive	0.6092	0.5564	0.4861	0.4645	0.3884

frequency-tuned method (FT) [9]. All baseline algorithms use default parameter settings given by the authors. Each image is resized to 256*256 pixels and regularly partitioned into 8*8 patches. Then features are extracted from each patch. Since our method does not utilize segmentation, the image segmentation based methods are not chosen for comparison.

To evaluate the effectiveness of our model in combing bottom-up features and top-down priors (Section III-D), three algorithms that also integrate bottom-up and top-down priors are used for comparison.

- *SMVJ*: Face detection is considered as a semantic cue and added into the bottom-up model [34]. (The code source is available at the page http://www.klab.caltech.edu/∼moran/fifadb/)
- *OptW*: Optimal combination weights for each feature map generated individually are learned from training images in a supervised manner [22].
- *SVM*: Several low, middle and high level image features are used to train a support vector machine (SVM) classifier that could predict salient regions [17]. The ratio of negative to positive samples is 1 for all the datasets.

3) Evaluation Metrics: All the algorithms are evaluated based on the following widely-used criteria. The receiver operator characteristic (ROC) is used to evaluate the similarity between the predicted and the ground-truth saliency maps. The ROC curve is obtained by trying all possible threshold values, and for each value, plotting the true positives rate (TPR) on the Y-axis against the false positive rate (FPR) value on the X-axis. For the convenience of evaluation, the area under ROC curve, denoted as AUC, is used to evaluate

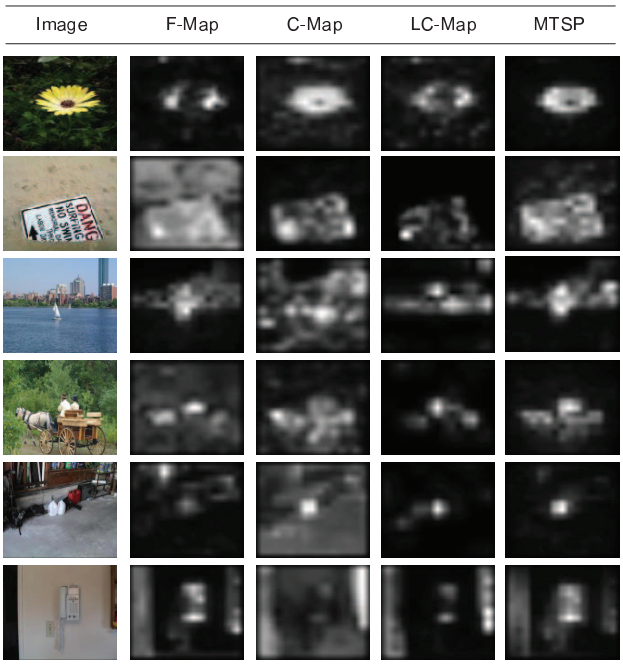


Fig. 7. Some examples for showing the advantages of integrating multiple features by MTSP. From left to right: the input image; the saliency map obtained from the feature of local energy; the saliency map obtained from the feature of color; the saliency map obtained from local contrast; the saliency map produced by MTSP that integrates together all types of features, including the local energy, color and local contrast.

the performance of various algorithms. We also compute the correlation coefficients (CC) [47] between he ground truth map and the predicted saliency map for evaluation. Let $M_g(x)$ and $M_p(x)$ (x generally denotes a position of a map) respectively be the ground-truth map and the predicted saliency maps, CC is defined as follows:

$$CC = \frac{\sum_x (M_g(x) - \mu_g)(M_p(x) - \mu_p)}{\sqrt{\sum_x (M_g(x) - \mu_g)^2 \sum_x (M_p(x) - \mu_p)^2}},$$

where μ_g and μ_p are the mean values of the two maps $M_g(x)$ and $M_p(x)$, and x indexes the pixels in the two maps.

4) Feature Extraction: The main target of our MTSP model is to integrate multiple visual features for joint saliency detection. According to the survey in [17], which summarizes and evaluates various features for saliency detection, we

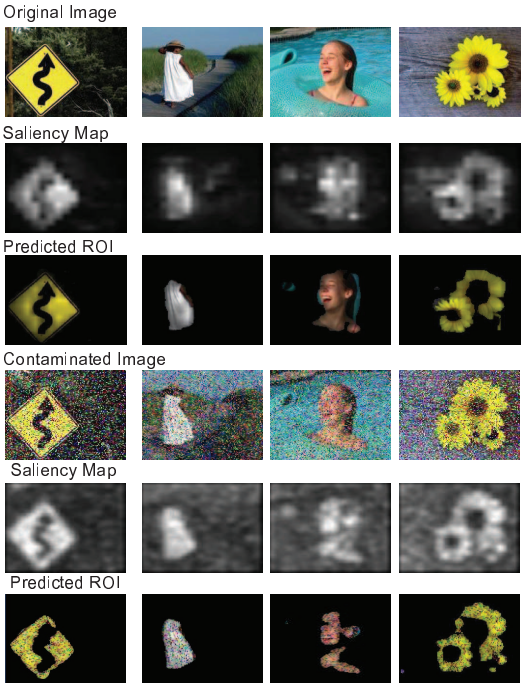


Fig. 8. Examples of using MTSP to detect saliency in images contaminated by noise.

choose three effective features [2]: *color*, *local energy* [45] and *local contrast* [2].

Color: We construct 6D color features by concatenating the values of Red, Green and Blue channels and their histograms: for the i-th pixel with RGB values (R_i, G_i, B_i), its corresponding feature vector is computed as $\{R_i, G_i, B_i, \mathcal{H}(R_i), \mathcal{H}(G_i), \mathcal{H}(B_i)\}$, where $\mathcal{H}(R_i) = -\log(Pr(R_i))$ (resp. $\mathcal{H}(G_i)$ and $\mathcal{H}(B_i)$) with $Pr(\cdot)$ being the estimated probability of a pixel value. In our experiments, the estimation of the probabilities is done by using 100 bins, and the feature vector of a patch is obtained by averaging over all the pixels in the patch.

Local energy: The steerable filter decomposition [45] provides a finer frequency decomposition that more closely corresponds to human visual processing. The basis functions of the steerable pyramid are directional derivative operators, that come in different scales and orientations. In the experiments, we use 3-scale steerable filter decomposition with 4 orientations. At each scale and orientation, the image is convolved by using the corresponding filter and decomposed into two parts, namely a low-pass part and high-pass part. The low-pass part is further processed by using the next filter at another scale and orientation. In this way, 12D features are produced by 3×4 high-pass parts and one feature is given by the final low-pass part, resulting in 13D feature vectors.

Local contrast: The local contrast is represented as the three channels corresponding to these conspicuity maps. Similar to [2], we calculate the three conspicuity maps for the color, intensity, and orientation contrasts. Each conspicuity map is obtained by center-surround difference [2] between a "center" fine scale and a "surround" coarser scale.

5) On Choosing the Parameter λ: The trade-off parameter λ in (4) has notable influence on the detection performance. Generally, the choice of this parameter depends on priors about the area of salient target. The smaller the λ is, the more patches may be identified as salient target, as shown in Figure 4. Nevertheless, the proposed method (MTSP) could work well under a range of parameter settings, as shown in Figure 5. While λ is ranging from 0.005 to 0.05, the ROC area is varying from 0.8075 to 0.8144, which demonstrates the model is insensitive to the parameter λ . For all experiments, we set $\lambda = 0.01$.

B. Results and Analysis

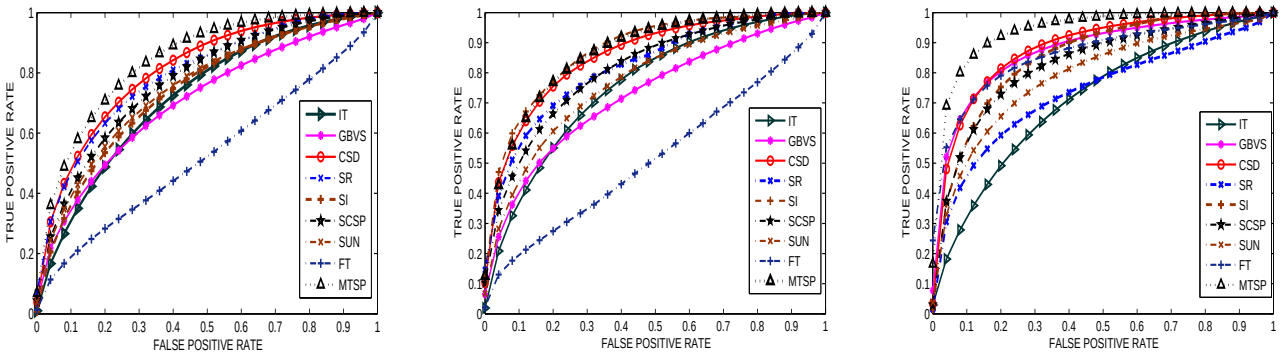
In this subsection, we systematically evaluate the proposed model on the saliency detection task. The evaluation consists of three aspects: 1) Investigating the performance improvement brought by integrating multiple features for detecting saliency collaboratively. 2) Comparing MTSP to the state-of-the-art bottom-up saliency algorithms. 3) Examining the effects of incorporating the top-down priors.

1) Advantages of Joint Inference: We first evaluate the proposed MTSP model performance (Section III-B2) for integrating multiple types of features to collaboratively produce saliency map. In order to perform saliency detection by utilizing multiple features, we consider a naive approach (denoted as "Naive") that computes an average map of the individual maps learnt from each feature by using the single-feature model presented in Section III-B1. For comparison, we also consider the performance of individual features based on the single-feature model in Section III-B1, resulting in three baselines: C-Map (color), F-Map (local energy) and LC-Map (local contrast).

Figure 6 shows the comparison results of the ROC curves on three datasets. It can be seen that the proposed model leads to better performance than the individually inferred or naively combined methods. Figure 7 presents some examples of the produced saliency maps, where the brighter areas correspond to the more salient regions of the image. It can be seen that the proposed MTSP performs reliably, while the approaches based on a single feature may fail sometimes. Most of our performance is gained from the consistent sparse matrix generally learned for multiple feature spaces. In summary, these results well verify the advantages of our formulation (4) for integrating the information of multiple visual features.

We also apply MTSP on the images contaminated by noise. For each image, its pixel values are corrupted by additive Gaussian noise with zero mean and standard deviation σ, where σ ranges from 0.05 to 0.35. Table I shows that MTSP perform much better than the single-feature model for handling noise. Figure 8 further provides some results by applying MTSP for detecting salient targets on the images contaminated by Gaussian noise. These results illustrate that the mechanism of joint inference (multi-task) is more robust to noise than the approach based on individual inference (single-task).

[2]Source codes of these features are available at http://people.csail.mit.edu //tjudd//WherePeopleLook//index.html.



(a) MIT Dataset (b) Bruce Dataset (c) MSRA Dataset

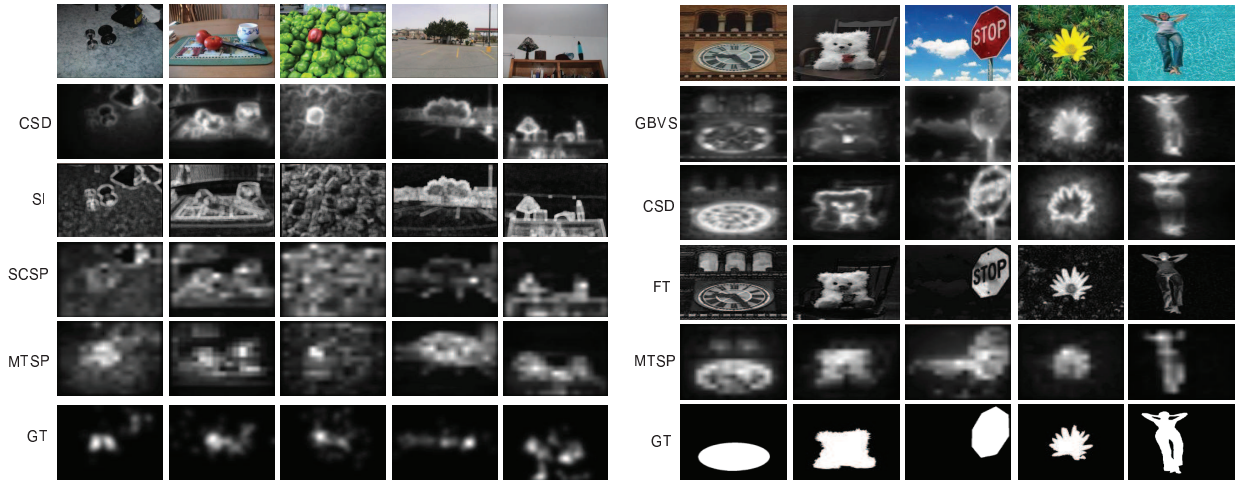
Fig. 9. The ROC curves of the proposed MTSP model and other eight state-of-the-art methods on MIT (a), Bruce (b) and MSRA (c) datasets.

TABLE II
THE AUC AND CC (CORRELATION COEFFICIENT) COMPARISON ON THE MIT, BRUCE AND MSRA DATASETS.

Criteria	Datasets	IT	GBVS	CSD	SR	SI	SCSP	SUN	FT	MTSP
AUC	MIT	0.7050	0.7234	0.8123	0.7936	0.7418	0.7513	0.7156	0.5265	**0.8462**
	Bruce	0.7291	0.7592	0.8594	0.8227	**0.8767**	0.8138	0.7745	0.5208	0.8708
	MSRA	0.7137	0.8765	0.8848	0.7377	0.8566	0.8314	0.7919	0.8653	**0.9247**
CC	MIT	0.2177	0.2497	0.3196	0.2793	0.2418	0.2813	0.2551	0.1435	**0.3467**
	Bruce	0.2873	0.3205	**0.4489**	0.3753	0.4337	0.3866	0.3264	0.1966	0.4422
	MSRA	0.3284	0.5742	0.5615	0.3232	0.4975	0.4972	0.4543	0.5685	**0.7044**

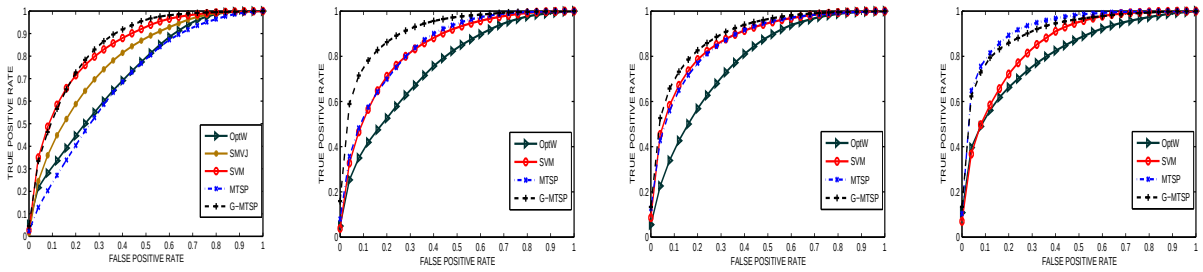
TABLE III
TO COMPENSATE THE CENTER BIAS EFFECT FOR HUMAN FIXATION DATA, THE MAUC COMPARISON IS ALSO PERFORMED ON THE MIT AND BRUCE DATASETS.

Dataset	IT	GBVS	CSD	SR	SI	SCSP	SUN	FT	MTSP
MIT	0.6161	0.6455	0.6518	0.6506	0.6258	0.6226	0.6349	0.5279	**0.6603**
Bruce	0.6307	0.6613	**0.6725**	0.6554	0.6696	0.6574	0.6547	0.5240	0.6721



(a) Bruce Dataset (b) MSRA Dataset

Fig. 10. (a) Some examples from the Bruce dataset. The rows from top to bottom are: original images, saliency maps produced by CSD, SI, SCSP, and MTSP, respectively. The last row is the human fixation map as ground truth. Note here that we only consider CSD, SI and SCSP, since the other competing algorithms are outperformed by our MTSP distinctly (see Figure 9).(b) Saliency results comparison on images from MSRA dataset. The rows for top to bottom are: original images, saliency maps produced by GBVS, CSD, FT, and MTSP, respectively. The last row is the ground truth. Note here that we do not consider the methods (e.g, IT) which have been outperformed by our MTSP greatly (see Figure 9).



(a) FiFa Dataset (b) MIT Dataset (c) Bruce Dataset (d) MSRA Dataset

Fig. 11. The ROC curves of the proposed G-MTSP and state-of-the-art methods on the four datasets.

2) Comparison to state-of-the-art methods: To further evaluate the competitiveness of the proposed model, we compare its results with the ones from existing eight state-of-the-art methods (Section IV-A2) on bottom-up saliency detection. Figure 9 and Table II show the comparison results. On the MIT, Bruce and MSRA datasets, especially the MSRA dataset, the proposed MTSP outperforms the other competing methods. Figure 10 shows some examples. It can be seen that MTSP exhibits stronger consistence with human eye fixations. These results illustrate that our MTSP is a competitive tool for saliency detection. The effectiveness of MTSP is mainly due to its ability of capturing information from multiple features in a unified inference procedure. This is achieved by seeking the consistent sparse components in order to encourage different features to produce consistent saliency maps. In this way, the cross-feature information is naturally modeled such that the multiple features will take effects collaboratively. And we emphasize that the proposed MTSP does not need to segment objects and no learning process from labeled images is taken into account.

As discussed in [46], [48], generally, for eye tracking dataset, there is a bias toward making early fixations near the center of an image, known as center bias [48]. To avoid this issue, we further adopt the evaluation framework described by [49], which involves modified ROC metrics for computing the area under the ROC curve, denoted as mAUC. Instead of using the non-fixed regions as negative samples, in mAUC the false positive fixations are fixations of a different image. For a given image, we choose all the other images in the dataset and calculate their averaged fixations as final false positive fixations for the given image. The mAUC results on the MIT and Bruce datasets are reported in TableIII, which shows that the performance of MTSP is competitive. Noting here that the mAUC measure generally underestimates the model performance if the ground truth saliency has central bias by itself, as pointed out by [48], [50]. Nevertheless, it is still a useful measure as a lower bound assessment of prediction ability of saliency detection model.

One may have noticed that MTSP does not outperform the most effective methods (SI and CSD) on the Bruce dataset. To explore the underlying reason of this phenomena, like [17], we analyze the consistency of human fixations over an image by measuring the *entropy* of the smoothed fixation map across viewers. Images with low entropy tend to contain one central object while images with high entropy are often rich in several different textures. We note that the images in Bruce dataset have relatively higher entropy: the MIT dataset has an average entropy of 0.10 and a standard deviation of 0.04, while the average entropy of the Bruce dataset is 0.15 (the standard deviation is 0.07). Also, as shown in Figure 12, MTSP tends to perform better on the images with lower entropy. This is the reason why the MTSP performs less improvements on the Bruce dataset than the other two datasets.

TABLE IV
THE AUC AND CC (CORRELATION COEFFICIENT) COMPARISON ON THE FOUR DATASETS.

Criteria	Dataset	SMVJ	SVM	OptW	MTSP	G-MTSP
AUC	FiFa	0.7953	0.8407	0.7517	0.6884	**0.8512**
	MIT	–	0.8494	0.7443	0.8462	**0.9063**
	Bruce	–	0.8752	0.7819	0.8708	**0.8989**
	MSRA	–	0.8582	0.8277	**0.9247**	0.9165
CC	FiFa	0.2623	0.3089	0.2556	0.1431	**0.4164**
	MIT	–	0.3316	0.2344	0.3467	**0.4535**
	Bruce	–	0.4525	0.2890	0.4422	**0.4626**
	MSRA	–	0.4578	0.4651	**0.7044**	0.6799

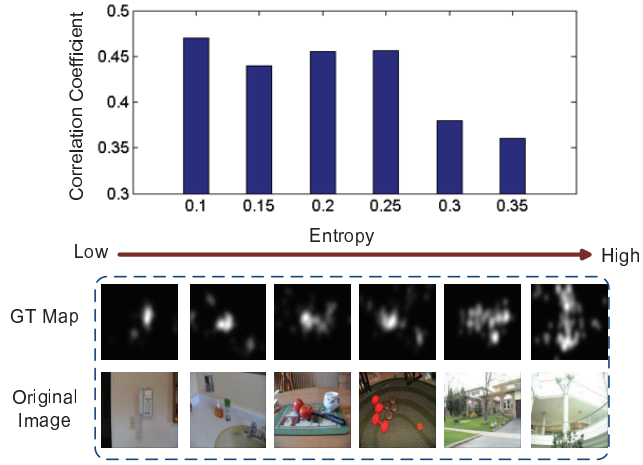


Fig. 12. Analysis of performance with respect to the entropy of human fixation map. We plot the performance of MTSP (in terms of CC) as a function of the entropy. It can be seen that MTSP achieves better performance on the images with lower entropy. These results are collected from the Bruce and MIT datasets.

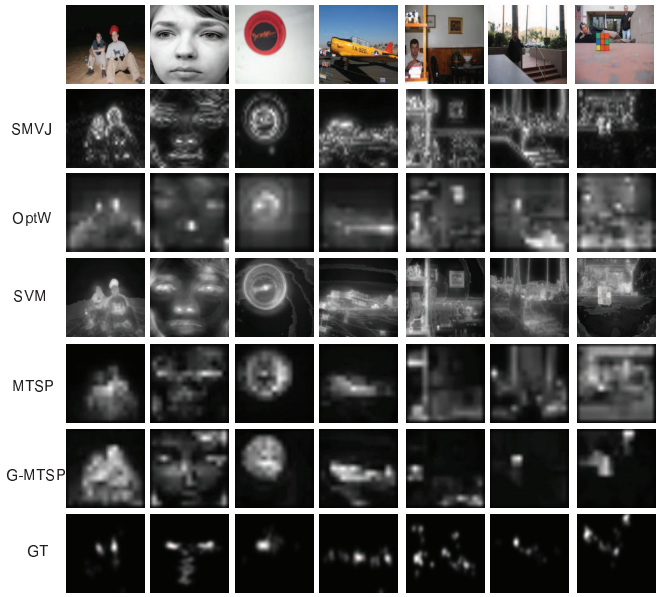


Fig. 13. Examples of the saliency maps produced by G-MTSP and four competing methods. Note here that MTSP is a pure bottom-up method, while all the others use top-down information.

3) Results of Incorporating Top-down Priors: As discussed in section III-D, MTSP can be generalized to incorporate the top-down priors represented by a label vector. To evaluate the effectiveness of G-MTSP, we use the FiFa, MIT, Bruce and MSRA datasets for experiments. Since most images of the FiFa dataset contain faces, similar to the [34], we use face detector to get face mask for each image as top-down priors. For the other three datasets, we learn an initial saliency map (i.e., a label vector) from the labeled images by using the method based on SVM [17].

Figure 11 and Table IV show the evaluation results in term of ROC and CC. Figure 13 exemplifies the performance of various algorithms. It can be seen that G-MTSP consistently outperforms the compared methods. In particular, G-MTSP can achieve much better than MTSP (which is a pure bottom-up model) on three datasets (FiFa, MIT and Bruce datasets). This result well verifies the effectiveness of incorporating top-down priors. As shown in Figure 11, G-MTSP significantly outperforms SVM on the FiFa, MIT and Bruce datasets. Since the top-down priors used by G-MTSP are exactly produced by SVM, this result illustrates that the integration of both bottom-up and top-down information is better than using top-down priors only.

One may have noticed that G-MTSP does not gain any improvement over MTSP on the MSRA dataset. The reason lies on the following two facts. On one hand, as can be seen from Table II and Table IV, SVM is outperformed by several unsupervised algorithms on this dataset. On the other hand, our unsupervised MTSP method has already achieved good performance, and so there may be little room left for better performance.

V. Conclusion

This paper introduced Multi-Task Sparsity Pursuit (MTSP), which is a generic model for saliency detection. First, we proposed that the recently established LRR [36] approach can fit well the single-feature based saliency detection. Second, we established a generalized formulation (4), so called MTSP, for integrating multiple visual features. Finally, we further generalize MTSP to incorporate the top-down priors encoded by a label vector, resulting in a general method that can handle the saliency detection problem under various settings: single-feature case, multi-feature case and the combination of top-down and bottom-up information. Experimental results well verified the effectiveness of the proposed method.

A key aspect of MTSP is its ability of integrating multiple visual features. In contrast with existing multi-feature based methods, MTSP integrates the information of multiple features into a unified inference procedure, which can be efficiently performed by solving a convex optimization problem. The proposed method seamlessly integrates multiple features to jointly produce the saliency map within a single inference step, and thus produces more accurate and reliable results. The proposed method may have general appealing for multi-task learning.

Acknowledgment

This work is partially support by project grant NRF2007IDM-IDM002-069 on Life Spaces from the IDM Project Office, Media Development Authority of Singapore, National Nature Science Foundation of China (60803072, 61100142, 90820013, 61033013, 61005030), Beijing Jiaotong University Science Foundation No. 2011JBM219, and 973 Program under Grant No.2007CB311002.

References

[1] M. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," *IEEE Transactions on Image Processing*, vol. 17, no. 5, pp. 633–644, 2008.

[2] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[3] T. Kadir and M. Brady, "Saliency, scale and image description." *International Journal of Computer Vision*, vol. 45, no. 2, pp. 1573–1405, 2001.

[4] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition - a gentle way," in *Workshop on Biologically Motivated Computer Vision*, 2002.

[5] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen, "Gaze-based interaction for semi-automatic photo cropping," in *SIGCHI conference on Human Factors in computing systems*, 2006, pp. 771–780.

[6] X. Hou and L. Zhang, "Thumbnail generation based on global saliency," in *International Conference on Cognitive Neurodynamics*, 2007.

[7] A. Bradley and F. Stentiford, "Visual attention for region of interest coding in jpeg 2000," *Journal of Visual Communication and Image Representation*, vol. 14, no. 3, pp. 232–250, 2003.

[8] Y. H. C. Rother, L. Bordeaux and A. Blake, "Autocollage," in *ACM Special Interest Group on GRAPHics and Interactive Techniques*, 2006, pp. 847–852.

[9] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE conference on computer vision and pattern recognition*, 2009, pp. 1597–1604.

[10] T. Avraham and M. Lindenbaum, "Esaliency (extended saliency): Meaningful attention using stochastic image modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 4, pp. 693–708, 2010.

[11] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2006, pp. 155–162.

[12] L. Elazary and L. Itti, "Interesting objects in natural scenes are more salient," in *Vision Science Society Annual Meeting*, 2007.

[13] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–6.

[14] S. Goferman, L. Zelnikmanor, and A. Tal, "Context aware saliency detection," in *IEEE conference on computer vision and pattern recognition*, 2010, pp. 2376–2383.

[15] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2006, pp. 545–552.

[16] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[17] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision*, 2009.

[18] J. Tsotsos, S. Culhane, W. Wai, Y. Lai, and N. Davis, "Modelling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1, pp. 507–545, 1995.

[19] V. Gopalakrishnan, Y. Hu, and D. Rajan, "Random walks on graphs for salient object detection in images," *IEEE Transactions on Image Processing*, vol. 19, no. 12, pp. 3232–3242, 2009.

[20] Y. Hu, D. Rajan, and L. Chia, "Detection of visual attention regions in images using robust subspace analysis," *Journal of Visual Communication and Image Representation*, vol. 19, no. 3, pp. 199–216, 2008.

[21] L. Itti and C. Koch, "A comparison of feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, pp. 473–482, 1999.

[22] L. Itti and C. Koch, "Feature combination strategies for saliency-based visual attention systems," *Journal of Electronic Imaging*, vol. 10, no. 1, pp. 161–169, 2001.

[23] V. Navalpakkam and L. Itti, "Search goal tunes visual features optimally," *Neuron*, vol. 53, pp. 605–617, 2007.

[24] Z. Lin, M. Chen, L. Wu, and Y. Ma, "The augmented Lagrange multiplier method for exact recovery of corrupted low-rank matrices," UIUC Technical Report UILU-ENG-09-2215, Tech. Rep., 2009.

[25] O. Meur, P. Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, 2006.

[26] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[27] O. Meur and J. Chevet, "Relevance of a feed-forward model of visual attention for goal-oriented and free-viewing tasks," *IEEE Transactions on Image Processing*, vol. 19, no. 11, pp. 2801–2813, 2010.

[28] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *IEEE conference on computer vision and pattern recognition*, 2007, pp. 1–8.

[29] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems*, 2009.

[30] J. Y. M. Zhu, H. Liu, and Y. Liu, "Visual saliency detection via sparsity pursuit," *IEEE Signal Processing Letters*, vol. 17, no. 8, pp. 739–742, 2010.

[31] M. Fazel, "Matrix rank minimization with applications," *PhD thesis*, 2002.

[32] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *IEEE conference on computer vision and pattern recognition*, 2007, pp. 1–8.

[33] A. Torralba, A. Oliva, M. Castelhano, and J. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features on object search," *Psychological Review*, vol. 113, no. 4, pp. 766–786, 2006.

[34] M. Cerf, J. Harel, W. Einhauser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems*, 2008, pp. 241–248.

[35] R. Peters and L. Itti, "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention," in *IEEE conference on computer vision and pattern recognition*, 2007.

[36] G. Liu, Z. Lin, and Y. Yu, "Robust subspace segmentation by low-rank representation," in *International Conference on Machine Learning*, 2010, pp. 663–670.

[37] J.-F. Cai, E. Candes, and Z. Shen, "A singular value thresholding algorithm for matrix completion," 2008.

[38] D. Bertsekas, *Constrained Optimization and Lagrange Multiplier Methods*, 1982.

[39] M. Afonso, J. Bioucas-Dias, and M. Figueiredo, "An augmented lagrangian approach to linear inverse problems with compound regularization," in *IEEE International Conference on Image Processing*, 2010, pp. 4169–4172.

[40] Y. Zhang, "Recent advances in alternating direction methods: Practice and theory," *Tutorial*, 2010.

[41] J. Eckstein and D. Bertsekas, "On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators," *Mathematical Programming*, vol. 55, pp. 293–318, 1992.

[42] G. Liu, Z. Lin, S. Yan, J. Sun, Y. Yu, and Y. Ma, "Robust recovery of subspace structures by low-rank representation," *Submitted to IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[43] J. Li, Y. Tian, T. Huang, and W. Gao, "Probabilistic multi-task learning for visual saliency in video," *International Journal of Computer Vision*, vol. 90, no. 2, pp. 150–165, 2010.

[44] V. Navalpakkam and L. Itti, "A goal oriented attention guidance model," *Lecture Notes in Computer Science*, vol. 2525, pp. 453–461, 2002.

[45] E. Simoncelli and W. Freeman, "A flexible architecture for multi-scale derivative computation," in *IEEE International Conference on Image Processing*, 1995, pp. 444–447.

[46] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A bayesian framework for saliency using natural statistics." *Journal of Vision*, vol. 8, no. 7, pp. 1–20, 2008.

[47] N. Ouerhani, R. Wartburg, and H. Hugli, "Empirical validation of the saliency-based model of visual attention," *Electronic Letters on Computer Vision and Image Analysis*, vol. 3, no. 1, pp. 13–24, 2004.

[48] B. Tatler, R. Baddeley, and I. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision Research*, vol. 45, no. 5, pp. 643–659, 2005.

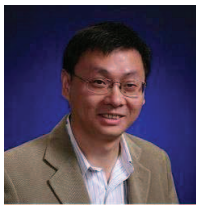
[49] N. Murray, M. Vanrell, X. Otazu, and C. Párraga, "Saliency estimation using a non-parametric low-level vision model," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[50] R. Carmi and L. Itti, "The role of memory in guiding attention during natural vision," *Journal of Vision*, vol. 6, no. 9, pp. 898–914, 2006.

Congyan Lang is currently an Associate Professor in the School of Computer and Information Technology, Beijing Jiaotong University, Beijing. She received her Ph.D from Beijing Jiaotong University in 2006. Her research interests include multimedia information retrieval and analysis, machine learning, and computer vision.

Guangcan Liu Dr. Liu received the bachelor degree of mathematics from Shanghai Jiao Tong University (SJTU) in 2004. Between 2006 and 2009, he was a visiting student at Visual Computing Group, Microsoft Research Asia. In 2010, he received the Ph.D. degree of computer science and engineering from SJTU. He is currently a postdoctoral research fellow at the department of electrical and computer engineering, National University of Singapore. His research interests include machine learning and computer vision. He is a member of the IEEE.

Jian Yu Dr. Yu received the B.S. degree in Applied Mathematics, M.S. degree in Mathematics, and Ph.D. degree in Applied Mathematics from Peking University, Beijing, P.R.China, in 1991, 1994 and 2000 respectively. In 2000, he joined the faculty member of Northern Jiaotong University. Since 2004, he became a full Professor and Head of Dept. of Computer ScienceBeijing Jiaotong University (previously named Northern Jiaotong University). At present, He is Vice Chair of Technical Committee of Artificial Intelligence and Pattern Recognition of China Computer Federation. His current research interests include machine learning, pattern recognition, and data mining, etc.



Shuicheng Yan Dr. Yan Shuicheng (M'06-SM'09) received the Ph.D. degree from the School of Mathematical Sciences,Peking University, in 2004. He spent three years as Postdoctoral Fellow at the Chinese University of Hong Kong and then at the University of Illinois at Urbana-Champaign, Urbana. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at National University of Singapore, and the founding lead of the Learning and Vision Research Group (http://www.lv-nus.org). Dr. Yan's research areas include computer vision, multimedia and machine learning, and he has authored or co-authored over 200 technical papers over a wide range of research topics. He is an associate editor of IEEE Transactions on Circuits and Systems for Video Technology, and has been serving as the guest editor of the special issues for TMM and CVIU. He received the Best Paper Awards from ACM MM'10, ICME'10 and ICIMCS'09, the winner prize of the classification task in PASCAL VOC'10, the honorable mention prize of the detection task in PASCAL VOC'10, 2010 TCSVT Best Associate Editor (BAE) Award, and the co-author of the best student paper awards of PREMIA'09 and PREMIA'11.